A COMPUTATIONAL METHOD FOR THE IDENTIFICATION OF CANDIDATE PROTEINS USEFUL AS ANTI-INFECTIVES

Field of Invention

The present invention provides a novel method for the identification of candidate proteins in pathogens useful as anti-infectives. More particularly, the present invention relates to candidate genes for these proteins. The invention further provides new leads for development of candidate genes, and their encoded proteins, and the study of their functional relevance to predictive, preventive or curative approaches. This computational method involves calculation of several sequence attributes and their subsequent analysis, leading to the identification of some outlier proteins in different pathogens. Thus, the present invention is useful for identification of some of the outlier proteins in pathogenic organisms. These outlier proteins are either virulence proteins or antigens or they may be used as drug targets. The outlier proteins from different genomes constitute a set of candidates for functional characterization through targeted gene disruption, microarrays and proteomics. Further, these proteins constitute a set of candidates for further testing in development of anti-infectives such as vaccine candidates, diagnostics or drug targets. Also, the genes encoding the candidate proteins are provided.

Background of the Invention and Prior Art Discussion

The progress in genome sequencing projects has generated a large number of inferred protein sequences from different organisms and, it is likely to increase in the coming years. The availability of complete genome sequences offers an opportunity for increased understanding of the biology of these organisms because it not only provides

biological insights on any given organism, but also provides substantially more information on the physiology and evolution of microbial species through comparative analysis (Fraser et al. 2000). The set of microbes whose genomes have been sequenced so far is a diverse one, ranging from organisms living under extreme condition of environment to model organisms of biology, and to some of the most important human pathogens (*see,* U.S. National Center for Biotechnology Information, U.S. National Institutes of Health, website).

It is expected that the availability of the information on the complete set of proteins from the infectious human pathogens will enable us to develop novel drugs to combat them. This is important in cases such as the emerging epidemic of multiple drug-resistant Mycobacterial isolates (Barry et al. 2000) although, so far, no new drugs derived from genomics-based discovery have been reported to be in a development pipeline (Black and Hare 2000). A paradigm for exploiting the genome to inform the development of novel antituberculars has been proposed, utilizing the techniques of differential gene expression as monitored by DNA microarrays coupled with the emerging discipline of combinatorial chemistry (Barry et al. 2000).

The whole genome sequences of microbial pathogens also present new opportunities for clinical applications such as diagnostics and vaccines (Weinstock et al. 2000). However, the predicted number of proteins encoded in different genomes is fairly large, and about half of that in any given genome is of unknown biological function (Fraser et al. 2000). Some of them are also unique in each organism. In this scenario, development of data mining tools and their application to decipher useful patterns in the protein sequence dataset can be useful for suitable experiments such as differential gene

expression, heterologous expression for large-scale (Weinstock et al 2000) and proteomics studies (Chakravarti 2000). Recently, it has been demonstrated that utilization of genome sequences by application of bioinformatics through genomics and proteomics can expedite the vaccine discovery process by rapidly providing a set of potential candidates for further testing (Chakravarti 2000 (a) and (b)). Presently data mining is being carried out using traditional computer programs that perform motif search or identify distinct domains differing in physico-chemical properties such as hydrophobicity, sequence conservation. The drawback of these methods is that the functions of a half to one third number of the proteins remain unknown even after their applications. Therefore, through the application of the presently available computation tools it is likely that potential new candidate for vaccines, diagnostics or drug targets are missed. Therefore, need exists for development of a computational tool that uses different sequence attributes of protein sequences instead of sequence patterns. Through such a shift in framework, the applicants have overcome this limitation. The novelty of the present invention is in development of method based on different attributes of protein sequences, which is useful for prediction of functional role in virulence, immuno-pathogenicity and drug-response.

References may be made to:

Barry, C.E. 3rd, Slayden, R.A., Sampson, A.E., and Lee, R.E. (2000). Use of genomics and combinatorial chemistry in the development of new antimycobacterial drugs. *Biochem. Pharmacol.* 59(3):221-31.

Black, T., and Hare, R. (2000). Will genomics revolutionize antimicrobial drug discovery? *Curr. Opin. Microbiol.* 3(5):522-7.

Chakravarti, D.N., Fiske, M.J., Fletcher, L.D., and Zagursky, R.J. (2000a). Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. *Vaccine* 19:601-12.

Chakravarti, D.N., Fiske, M.J., Fletcher, L.D., and Zagursky, R.J. (2000b). Mining genomes and mapping proteomes: identification and characterization of protein subunit vaccines. *Dev. Biol.* (Basel) 103:81-90.

Fauchere, J.L. and Pliska, V. (1983). Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino acid amides. *Eur. J. Med. Chem. -Chim. Ther.* 18:369-375.

Fraser, C.M., Eisen, J., Fleischmann, R.D., Ketchum, K.A. and Peterson, S. (2000). Comparative genomics and understanding of microbial biology. *Emerging Infectious Diseases*, Vol.6, 505-512.

Hopp, T.P. and Woods, K.R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA*, 78: 3824-3828.

Kester, K.E., McKinney, D.A., Tornieporth, N., Ockenhouse, C.F., Heppner, D.G., Hall, T., Krzych, U., Delchambre, M., Voss, G., Dowler, M.G., Palensky, J., Wittes, J, Cohen, J, and Ballou, W.R. (2001). Efficacy of Recombinant Circumsporozoite Protein Vaccine Regimens against Experimental Plasmodium falciparum Malaria. *J. Infect. Dis.* 183(4): 640-647.

Kyte, J. and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105-132.

Mande, S.S., Gupta, N., Ghosh, A. and Mande, S.C. (2000). Homology model of a novel xylanase: Molecular basis for high-thermostability and alkaline stability. *J. Biomol. Str. Dyn.* 18:137-144.

Mobley, H.L., Garner, R.M., Chippendale, G.R., Gilbert, J.V., Kane, A.V., and Plaut, A.G. (1999). Role of Hpn and NixA of Helicobacter pylori in susceptibility and resistance to bismuth and other metal ions. *Helicobacter* 4(3):162-169.

Nakashima, H. and Nishikawa, K. (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue pair frequencies. *J. Mol. Biol.* 238: 54-61.

Ramachandran, S., Nandi, T., Ghai, R., B-Rao, C., Brahmachari, S. K., and Dash, D. Analysis of complete genome sequences using a novel complexity measure (submitted).

Ramakrishnan, L., Federspiel, N.A., and Falkow, S. (2000). Granuloma-specific expression of Mycobacterium virulence proteins from the glycine-rich PE-PGRS family. *Science* 288(5470):1436-9.

Roland, L. and Eisenberg, D. (1992). Protein in Sequence Analysis Primer. Eds. Gribskov, M. and Devereux, J. Oxford University Press, 61-87.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, 229: 834-838.

Varadarajan, R., Nagarajaram, H.A., and Ramakrishnan, C. (1996). A procedure for the prediction of temperature-sensitive mutants of a globular protein based solely on the amino acid sequence. *Proc. Natl. Acad. Sci. USA* 93:13908-13.

Weinstock, G.M., Smajs, D., Hardham, J., and Norris, S.J. (2000). From microbial genome sequence to applications. *Res. Microbiol.* 151(2):151-8.

Wien Klin Wochenschr. (1997). Aug 8; 109(14-15):551-6. Comparative genomics of mycoplasmas.

Wootton, J.C. (1994). Non globular domains in protein sequences: Automated segmentation using complexity measures. *Comput. Chem.*, 18: 269-285.

Objects of the Invention

The main object of the present invention is to provide a computational method for identification of proteins useful as anti-infectives. These anti-infectives are vaccine candidates, diagnostics or drug targets.

Another object of invention is to provide proteins with unusual sequence characteristics identified as outliers in different pathogens.

Yet another object of the invention is for providing the use of gene sequences encoding the proteins useful as candidate anti-infectives.

Summary of the Invention

The present invention relates to a computational method for the identification of candidate proteins useful as anti-infectives. The invention particularly describes a novel strategy to identify outlier proteins in different genomes of pathogens. These anti-infectives are vaccine candidates, diagnostics or drug targets.

Brief Description of the Accompanying Drawings

Figure 1 represents the one of the bivariate relationship for *Mycobacterium tuberculosis*.

Detailed Description of the Invention

Accordingly, the present invention provides a novel computational method for the identification of candidate proteins in pathogens useful as anti-infectives. Computational algorithms based on general principles are used to carry out data mining to decipher useful patterns for sequence characterization and classification. This computational method involves calculation of several sequence attributes and their subsequent analysis, leading to the identification of some outlier proteins in different pathogens. Thus, the present invention is useful for identification of some of the outlier proteins in pathogenic organisms. These outlier proteins are either virulence proteins or antigens or used as drug targets. The outlier proteins from different genomes constitute a set of candidates for functional characterization through targeted gene disruption, microarrays and proteomics. Further, these proteins constitute a set of candidates for further testing in development of anti-infectives such as vaccine candidates, diagnostics or drug targets. Also, the genes encoding the candidate proteins are provided.

The invention provides a set of candidate proteins and genes for further evaluation as diagnostic or vaccine candidate or useful for testing in diagnostics or drug susceptibility for human pathogens. The method of the invention is based on the analysis of protein sequence attributes instead of sequence patterns linked to biochemical functions. The present method is independent of the discrepancy inherent with such an

approach. The invention provides a computational method, which involves multivariate analysis using Principle Component Analysis (PCA). The proteins termed 'outliers' were found to be excluded from the protein clusters in various pathogens' genomes. Several unique sequences were located on homology analyses of these 'outliers' protein sequences with those in Swiss Prot and PIR database. Some outlier sequences turned out to be identical or homologous to the virulent proteins implicated with antigenic and drug susceptible responses. By this approach, proteins could be identified (short-listed) for further testing in development of anti-infectives in pathogenic organisms.

Computational algorithms based on general principles are needed to carry out data mining to decipher useful patterns for sequence characterization and classification.

The invention has utility for providing new leads for development of anti-infectives of diagnostic, preventive and curative potential.

The present invention relates to a computational method for the identification of candidate proteins useful as anti-infectives.

Accordingly, the present invention provides a novel method for identifying the candidate proteins useful as anti-infectives, said method comprising:

i) calculating computationally the different sequence-based attributes from all the protein sequences of the selected pathogenic organisms.

ii) clustering computationally all the proteins of a genome based on these sequence-based attributes using Principle Component Analysis.

iii) identifying computationally the outlier proteins sequences which are excluded from the main cluster.

iv) matching the outlier protein sequences with the protein sequences in various databases.

v) selecting the unique outlier protein sequences not homologous to any of the protein sequences searched above.

vi) validating computationally the protein sequences as anti-infectives by comparing with the known protein sequences that are biochemically characterized in the pathogen genome.

In an embodiment of the present invention, the protein sequence data is taken from any organism, for example, but not limited to, organisms such as *Borrelia burgdorferi, Campylobacter jejuni, Chlamydia pneumoniae, Chlamydia trachomatis, Haemophilus influenzae, Helicobacter pylori, Leishmania major, Mycoplasma genetalium, Mycoplasma pneumoniae, Mycobacterium tuberculosis, Neisseria meningitis, Pseudomonas aeruginosa, Plasmodium falciparum, Rickettsia prowazekii, Treponema pallidum,* and *Vibrio cholerae.*

In another embodiment of the present invention, different sequence-based attributes used for identification of candidate anti-infective proteins are selected from the group comprising fixed protein and variable protein attributes.

In still another embodiment of the present invention, the fixed protein attributes are selected from the group comprising: percentage of charged amino acids, percentage hydrophobicity, distance of protein sequence from a fixed reference frame, measure of dipeptide complexity of protein, and measure of hydrophobic distance from a fixed reference frame.

In yet another embodiment of the present invention, the variable attribute is the distance of the protein sequence from a variable reference frame.

In one more embodiment of the present invention, the cluster analysis is carried out by Principle Analysis Technique using correlation coefficient between the attributes.

In one other embodiment of the present invention, the steps i) to iv) and vi) are performed computationally.

In an embodiment of the present invention, the clustering of the proteins is based upon analysis of sequence attributes instead of sequence pattern linked to biochemical functions.

In another embodiment of the present invention, the unique outlier protein sequences non-homologous to the known anti-infective sequences are specifically identified in the following pathogens, but not limited to, such as *B. burgdorferi, C. jejuni, C. pneumoniae, C. trachomatis, H. influenzae, H. pylori, L. major, M. genetalium, M. pneumoniae, M. tuberculosis, N. meningitis, P. aeruginosa, P. falciparum, R. prowazekii, T. pallidum*, and *V. cholerae.*

In still another embodiment of the present invention, the unique outlier sequences obtained by the method of invention that can serve as potential anti-infective candidates are listed in Table 1 and List 1.

In yet another embodiment of the present invention, the unique outlier hypothetical protein sequences from pathogenic genomes that can serve as anti-infective candidates are listed in Table 2.

In one more embodiment of the present invention are the genes encoding the unique proteins useful as anti-infectives.

Another embodiment of the present invention is the computer system, comprising a central processing unit, executing the DISTANCE program, followed by clustering of the protein sequences based on different attributes using by Principle Component Analysis, all stored in a memory device accessed by CPU, a display on which the central processing unit displays the screens of the above mentioned programs in response to user inputs; and a user interface device.

In an embodiment of the present invention are the unique outlier hypothetical protein sequences from pathogenic genomes that can be used for diagnostic purposes.

In another embodiment of the present invention are the unique outlier hypothetical protein sequences from pathogenic genomes that can be used as vaccine candidates.

In still another embodiment of the present invention are the unique outlier hypothetical protein sequences from pathogenic genomes that can be used for therapeutic purposes.

Unique outlier protein sequences non-homologous to the known anti-infective sequences are specifically identified in the following pathogens, but not limited to, such as *B. burgdorferi, C. jejuni, C. pneumoniae, C. trachomatis, H. influenzae, H. pylori, L. major, M. genetalium, M. pneumoniae, M. tuberculosis, N. meningitis, P. aeruginosa, P. falciparum, R. prowazekii, T. pallidum*, and *V. cholerae*.

Unique outlier protein sequences obtained by the method of invention that can serve as potential anti-infective candidates and having known properties are listed in Table 1 and List 1.

Unique outlier hypothetical protein sequences from pathogenic genomes that can serve as anti-infective candidates listed in Table 2. These protein sequences have hypothetical functions.

List 1 contains all the protein sequences that were marked as outlier by clustering method. These sequences were obtained from NCBI database.

Other and further aspects, features and advantages of the present invention will be apparent from the following description of the presently preferred embodiments of the invention given for the purpose of disclosure.

Description of tables and sequence lists

List 1 contains all the protein sequences that were marked as outliers by clustering method. These sequences were obtained from NCBI database (*see,* U.S. National Center for Biotechnology Information, U.S. National Institutes of Health, website).

Table 1 gives the list of outlier proteins with known functions.

Table 2 gives the list of outlier proteins with hypothetical functions.

Brief description of computer program:

The software program was written in PERL (Practical Extraction and Reporting Language) and operated on a Silicon Graphics Origin 200 using IRIX 6.5 operating system. The computer program gives a numerical data of the different attribute, column-

wise, for each protein in one record along with its GI number. The values in each column represent the values of the different variates in the multivariate analysis. Using the rationale described above we have developed the data mining software and a software copyright has been filed.

Statistical Analysis

All statistical procedures were carried out using the SAS package (SAS Institute Inc., USA). Principal Component Analysis using correlation coefficients between the variates was carried out using this package.

Sequence analysis

Homology analysis was carried out using the Wisconsin Package Version 10.0, Genetics Computer Group (GCG), Madison, Wisconsin.

Details of the Invention

The whole genome sequences of microbial pathogens present new opportunities for clinical applications such as diagnostics and vaccines (Weinstock et al. 2000). The present invention provides new leads for the development of candidate genes, and their encoded proteins, in view of their functional relevance to drug responses for use in predictive, preventive or curative approaches.

The protein sequences of several pathogens were obtained computationally from the existing databases (NCBI, genbank/genomes/bacteria). Different sequence attributes like hydrophobicity, charge and measures of compositional distance and dipeptide complexity by a specially developed computer program 'DISTANCE' was used for computation. The attribute profile was obtained for all the proteins for each of the

pathogenic genomes. These sequence-based attributes were then used to carry out cluster analysis by the Principal Component Analysis technique using correlation coefficients between the attributes. The proteins falling outside the protein cluster in each genome were identified and termed as outlier proteins. These outlier proteins were compared by BLAST with the sequence of known protein anti-infectives to identify potential candidates for anti-infective lead molecules which can be envisaged to be useful for predictive, preventive and curative purposes against pathogenic infections.

Accordingly, the invention provides a computer-based method for identifying the candidate proteins useful as anti-infectives, which comprises:

1. calculating computationally the different sequence-based attributes from all the protein sequences of the selected pathogenic organisms.

2. clustering computationally all the proteins of a genome based on these sequence-based attributes using Principle Component Analysis.

3. identifying computationally the outlier proteins sequences which are excluded from the main cluster.

4. matching the outlier protein sequences with protein sequences in various databases.

5. selecting the unique outlier protein sequences not homologous to any of the protein sequences searched above.

6. validating computationally the protein sequences as anti-infectives by comparing them with known protein sequences that are biochemically characterized in the pathogen genome.

In an embodiment of the invention, the protein sequence data may be taken from any organism, specifically, but not limited to, organisms such as *B. burgdorferi, C. jejuni, C. pneumoniae, C. trachomatis, H. influenzae, H. pylori, L. major, M. genetalium, M. pneumoniae, M. tuberculosis, N. meningitis, P. aeruginosa, P. falciparum, R. prowazekii, T. pallidum*, and *V. cholerae.*

In an embodiment, the non-homologous outlier protein sequence may be compared with that of known anti-infective sequences in the selected pathogens. Several unique outlier sequences were identified to be similar to sequences known to play a role in anti-infectives. These unique sequences obtained by the method of the invention can serve as potential anti- infective candidates.

In another embodiment of the present invention, different sequence-based attributes used for identification of candidate anti-infective proteins comprise charge, hydrophobicity, distance from fixed and variable point of reference, hydrophobic distance and dipeptide complexity.

In another embodiment, the attributes may be of fixed type or variable type.

In another embodiment of the invention, the computer system comprises a central processing unit, executing the DISTANCE program, followed by clustering of the protein sequences based on different attributes using by Principle Component Analysis, all stored in a memory device accessed by CPU, a display on which the central processing unit displays the screens of the above mentioned programs in response to user inputs; and a user interface device.

The particulars of the organisms such as their name, strain, accession number in NCBI database and other details are given below:

List 1.

Genomes	**Accession No.**	**No. of bp(s)**	**Date of completion**
B. burgdorferi	NC_001318	910724 bp	Dec. 17, 1997
C. jejuni	NC_002163	1641481 bp	Feb. 10, 2000
C. pneumoniae CWL029	NC_000922	1230230 bp	Dec. 1, 1998
C. trachomatis	NC_000117	1042519 bp	May 20, 1998
H. influenzae	NC_000907	1830138 bp	July 25, 1995
H. pylori	NC_000915	1667867 bp	Aug. 6, 1997
L. major		chromosome 1	
M. genetalium	NC_000908	580074 bp	Jan. 8, 2001
M. pneumoniae	NC_000912	816394 bp	June 15, 1996
M. tuberculosis	NC_000962	4411529 bp	June 11, 1998
N. meningitis MC58	NC_002183	2272351 bp	Feb. 25, 2000
P. aeruginosa	NC_002516	6264403 by bp	May 16, 2000
P. falciparum		chromosome 2, 3	
R. prowazekii	NC_000963	1111523 bp	Nov. 12, 1998
T. pallidum	NC_000919	1138011 bp	Mar. 6, 1998
V. cholerae	NC_002505	2961149 bp	June 14, 2000
	NC_002506	1072315 bp	June 14, 2000

Genomes	**Total number of proteins**
B. burgdorferi	850
C. jejuni	1634
C. pneumoniae	1052
C. trachomatis	894
H. influenzae	1709
H. pylori	1553
L. major	683
M. genetalium	467
M. pneumoniae	677
M. tuberculosis	3918
N. meningitis	2025
P. aeruginosa	5565
P. falciparum	422
R. prowazekii	834
T. pallidum	1031
V. cholerae	3828

Another embodiment of the invention is the use of the genes encoding the proteins identified by the methods of the invention.

The invention is further explained with the help of the following examples which are given by illustration and should not be construed to limit the scope of the present invention in any manner.

EXAMPLES

Example 1:

DISTANCE:

The purpose of the program is to computationally calculate various sequence-based attributes of the protein sequences.

The program works as follows:

The internet downloaded FASTA format files obtained from NCBI (*see,* U.S. National Center for Biotechnology Information, U.S. National Institutes of Health, website) were saved by the name <organism name>.faa and passed as input to the PERL program which computes the different attributes of protein sequences.

Input/Output format:

Downloaded Files and their format:

<organism name>.faa: file which stores the annotation and the protein sequence. <organism name> refers to

BB (*Borrelia burgdorferi*), BS (*Bacillus subtilis*), CJ (*Campylobacter jejuni*), CP (*Chlamydia pneumoniae*), CT (*Chlamydia trachomatis*), HI (*Haemophilus influenzae*), HP (*Helicobacter pylori*), LP (*Leishmania major*), MG (*Mycoplasma genetalium*), MP (*Mycoplasma pneumoniae*), MTUB (*Mycobacterium tuberculosis*), NM (*Neisseria meningitis*), PAER (*Pseudomonas aeruginosa*), PF (*Plasmodium falciparum*), RP (*Rickettsia prowazekii*), TP (*Treponema pallidum*), VCHO (*Vibrio cholerae*)

Format: FASTA

">gi|" <annotation>

<< the entire protein sequence.....

For example,

>gi|2314605|gb|AAD08472| histidine and glutamine-rich protein

MAHHEQQQQQANSQHHHHHHAHHHHYYGGEHHHHNAQQHAEQQAEQQAQ

QQQQQQAHQQQQQKAQQQNQQY (SEQ ID NO:14)

>gi|3261822|gnl|PID|e328405 PE_PGRS

MIGDGANGGPGQPGGPGGLLYGNGGHGGAGAAGQDRGAGNSAGLIGNGGAGG

AGGNGGIGGAGAPGGLGGDGGKGGFADEFTGGFAQGGRGGFGGNGNTGASGG

MGGAGGAGGAGGAGGLLIGDGGAGGAGGIGGAGGVGGGGGAGGTGGGGVAS

AFGGGNAFGGRGGDGGDGGDGGTGGAGGARGAGGAGGAGGWLSGHSGAHG

AMGSGGEGGAGGGGGARGEAGAGGGTSTGTNPGKAGAPGTQGDSGDPGPPG

(SEQ ID NO:18)

>gi|.....

The output file: <organism_name>.mdis

Format:

for example format of mtub.mdis:

Gene name	**Length (L)**	**% Hydrophobicity**	**% charge**	**D_{fixed}**	**$D_{var,high\ complexity}$**	**Dipeptide**	**D_{phobic}**
>gi|2808711|gnl|PID|e1245984	507	49.9	25.44	63.06	53.38	90	53.18
>gi|3261513|gnl|PID|e1299736	402	60.95	21.39	68.64	40.88	81	60.3
>gi|1552556|gnl|PID|e266921	385	58.18	27.27	71.16	43.13	79	59.25
>gi|1552557|gnl|PID|e266922	187	56.15	25.67	34.79	23.17	22	29.1
>gi|1552558|gnl|PID|e266923	714	51.12	27.87	87.22	80.66	154	77.04
>gi|1552559|gnl|PID|e266924	838	53.46	27.33	116.02	88.15	196	97.71
>gi|1552560|gnl|PID|e266925	304	61.84	17.11	54.21	34.79	49	47.55
>gi|.....							

Example 2:

Fixed Protein attributes:

We developed a framework for statistical analysis using the following attributes of proteins. The attributes used here are the hydrophobicity, charge, and different types of compositional characteristics of a protein. Each attribute was quantified using a measure and each measure uses a reference frame for computation defined later in this section.

The attributes were treated as variates in the statistical analysis. The variates were classified into two categories, namely, 'fixed' and 'variable.' In the case of 'fixed' variates, the reference frame for analysis of different organisms (genomes) is fixed. Thus the reference frame in these cases is not organism specific. For example, a particular scale of hydrophobicity is fixed for the analysis of protein sequences across all organisms. In the case of `variable' variates, the reference frame for analysis of different organisms (genomes) varies from one to another. In these cases, the reference frame is organism specific.

In this work, we have included variates with reference frames that are not organism specific and that are organism specific because our objective was to analyze the different characteristics of the proteins in one module to enable us to draw inferences with significance and practical utility. Thus, proteins falling as outliers based on all these variates have very different characteristics in general and also from the rest members of the genome.

L is the length of the protein in number of amino acids.

The group of charged amino acids, hydrophobicity scale used, expected number of occurrences of different amino acids, expected number of different dipeptides in a protein, expected number of hydrophobic amino acids - based on a particular hydrophobicity scale - each constitute a reference frame for the different measures used in this work. These measures are described below.

Fixed variates:

Variate 1: is the percent of charged amino acids in a given protein. The charged amino acids were Aspartic acid (D), Glutamic acid (E), Lysine (K) and Arginine (R).

% of Charge is given by

$$\frac{\text{Number of charged amino acids}}{L} \quad X\ 100 \qquad (1)$$

Variate 2: is the percent hydrophobicity of the protein. We have used several hydrophobic scales given by Fauchere & Pliska scale (Fauchere and Pliska, 1983), Hopp & Woods (1981), Kyte & Doolittle (1982) and Rose scale (Rose et al. 1985) to classify the amino acids into hydrophobic and hydrophilic groups respectively.

Percent Hydrophobicity is given by

$$\frac{\text{Number of hydrophobic amino acids}}{L} \quad X\ 100 \qquad (2)$$

Variate 3: is a measure of distance of a protein sequence from a fixed reference frame. The distance is measured according to the formula:

$$D_{fixed} = \sqrt{\sum_{x=1}^{20} (O_x - E_x)^2} \qquad (3)$$

O_x is the observed number of xth amino acid in the protein and E_x is the expected number of xth amino acid in the same protein. In this case, E_x is L/20 considering all amino acids to be uniformly distributed in the fixed reference frame. D_{fixed} /L is a normalized measure of distance for the protein.

Variate 4: is a measure of the dipeptide complexity of a protein. The reference frame here is the maximum number of dipeptides possible in the protein for its length. The measure is given by

(i) for proteins of L < 800 amino acids

$$\frac{\text{No. of different peptides observed in the protein}}{(L/2)} \qquad (4)$$

and

(ii) for proteins of L > 800 amino acids

$$\frac{\text{No. of different peptides observed in the protein}}{400} \qquad (5)$$

Variate 5: is a measure of hydrophobic distance of a protein in a genome from a fixed reference frame.

$$D_{phobic} = \sqrt{\sum_{x=1}^{20} (O_x - E_x)^2} \qquad (6)$$

O_x is the observed number of xth hydrophobic amino acid in the protein and E_x is the expected number of xth hydrophobic amino acid in the same protein. In this case,

$$E_x = \frac{\text{total no. of hydrophobic amino acids in the protein}}{z}$$

The computation of E_x assumes uniform distribution of the different hydrophobic amino acid types; z = the number of types of hydrophobic amino acids identified according to a particular hydrophobic scale. This is the fixed reference frame. z will vary according to the hydrophobic scale used. For example in the Kyte & Doolittle scale, z is 13; in the Hopp and Woods scale, z is 11; in the Fauchere & Pliska scale, z is 11; and in the Rose scale, z is 8. D_{phobic} /L is a normalized measure of hydrophobic distance of a protein.

Example 3:

Variable Protein Attributes:

Variate 6: is the distance of a protein sequence in a genome from a variable reference frame. In this case the distance Dvar, high complexity has the same formula as that in Variate 3 but Ex is calculated according to the formula:

$$\mathbf{E_x = f_x X L} \qquad (7)$$

where f_x is the frequency of occurrence of the xth amino acid in the set of proteins that are of 'high sequence complexity' within the same genome. For this purpose, we first run the protein sequences encoded in the genome through our sequence complexity analysis computer program (Ramachandran et al.) and classify the proteins into 2 sets, namely, 'high complexity' and 'low complexity' according to the fraction of the low complexity sequences present in each protein.

The frequency of each of the 20 amino acids from the high complexity set of proteins was computed by calculating the number of occurrences of the xth (x = 1 to 20) amino acid in the proteins set divided by the total number of amino acids in the same set.

The frequency of occurrences of different amino acids in this dataset is referred to as the variable reference frame because the frequency of the different amino acids appearing in the high complexity set of proteins are unequal to each other and varies from one genome to another. As in Variate 3, Dvar, high complexity / L is a normalized measure of distance with respect to the variable reference frame.

Example 4:

Clustering by Principle Component analysis

A representation of one of the bivariate relationship for *M. tuberculosis* is shown in Figure 1. The ellipse of confidence limit at 80% is also shown. The relationship between the variate D_{fixed}/L and % hydrophobicity shows that most of the proteins in different genomes cluster into a large dense group. A few proteins tend to fall outside the cluster in different organisms. Similar observations were made with all types of bivariate plots and with all organisms (data not shown). These observations indicate the clustering nature of the proteins from different organisms with respect to the protein attributes, and this feature could explain the nature of uniformity observed in the distribution patterns discussed in the previous section. The proteins that fall outside the clusters are termed as 'outliers' in this work. The number of outliers in different organisms vary from one organism to another.

In the present invention the most widely used hydrophobicity scales, charge composition, and various distance measures based on amino acid frequencies have been used. When one hydrophobicity scale is used instead of another, then the list of the outliers changes only very slightly. Most of the outlier proteins are common to all the 4

scales. We have included in our list all the outliers identified using all the 4 different scales of hydrophobicity each taken one at a time.

A comprehensive study has been done to identify the outliers in different genomes by principal components analysis at 0.8 of cumulative proportion of variance. The number of outliers identified in different genomes is given (Tables 1 & 2). It is evident that the number of outliers does not have a clear relationship with the total number of proteins encoded in the different genomes. This indicates that the properties of the outlier proteins do not follow a common trend with respect to the number of proteins encoded in a genome (or the genome size). The number of outliers in the case of *P. falciparum* and *L. major* is with respect to the partial genomic sequences. A clearer picture will emerge after the whole genome is sequenced and the protein coding regions are identified.

Example 5:

Prediction of anti-infective annotation in *M. tuberculosis*

Seven outlier sequences were identified in *M. tuberculosis* (Tables 1 & 2). Among these, three protein sequences correspond to glycine-rich protein PE_PGRS (Poly E rich proteins) of *M. tuberculosis*. The amino acid sequences of these can be retrieved from the NCBI database (*see*, U.S. National Center for Biotechnology Information, U.S. National Institutes of Health, website). The PE_PGRS proteins have been implicated in virulence in this pathogen (Ramakrishnan et al., 2000). These unique outlier protein sequences can therefore be predicted to be potential candidates for an anti-infective approach.

Example 6:

Prediction of anti-infective annotation in *H. pylori*

Eight outlier sequences were identified in *H. pylori* (Tables 1 & 2). Bacteria lacking one these outliers, i.e., a histidine rich protein, cultured *in vivo*, are more susceptible than is the wild type to bismuth and Ni^{2+} (Mobley et al 1999). These unique outlier protein sequences can therefore be predicted to be potential candidates for an anti-infective approach.

Example 7:

Prediction of anti-infective annotation in *P. falciparum*

Five outlier sequences were identified in *P. falciparum* (Tables 1 & 2). The circumsporozite protein was evaluated as a vaccine candidate (Kester et al 2001). These unique outlier protein sequences can therefore be predicted to be potential candidates for an anti-infective approach.

The particulars of the organisms such as their name, strain, accession number in NCBI database and other details are given above in List 1.

Table 1: List of proteins with known functions

Organism	GI Number	Protein function	SEQ ID NO:
Eubacteria			
CJ	6967728	highly acidic protein	SEQ ID NO: 1
	6969129	small hydrophobic protein	SEQ ID NO: 2
	6968493	putative coiled coil protein	SEQ ID NO: 3
	6968611	highly acidic protein	SEQ ID NO: 4
CP	4376663	histone like protein2	SEQ ID NO: 5
CT	3522902	hypothetical protein-possible frameshift with CT593	SEQ ID NO: 6
	3328438	histone like protein2	SEQ ID NO: 7
HI	1573353	tol A	SEQ ID NO: 8
	1574049	thiamin ABC transporter	SEQ ID NO: 9
	1574645	heme exporter protein B	SEQ ID NO: 10
	1573009	recombination protein	SEQ ID NO: 11
HP	2313421	poly E-rich protein	SEQ ID NO: 12
	2314604	histidine rich, metal binding polypeptide	SEQ ID NO: 13
	2314605	histidine and glutamine rich protein	SEQ ID NO: 14
MG	1046012	cytaadherence accessory protein	SEQ ID NO: 15
	1046097	cytaadherence accessory protein	SEQ ID NO: 16
MP	1674069	adhesin related protein	SEQ ID NO: 17
MTUB	3261822	PE_PGRS	SEQ ID NO: 18
	2894254	PE_PGRS	SEQ ID NO: 19
	2924449	PE_PGRS	SEQ ID NO: 20
	1781260	PPE	SEQ ID NO: 21
PAER	9947600	KdpF	SEQ ID NO: 22
	9951563	alginate regulatory protein A1gP	SEQ ID NO: 23
	9951352	PhaF	SEQ ID NO: 24
TP	3323280	dicarboxylate transporter	SEQ ID NO: 25
VCHO	9654609	iron (III) ABC transporter, permease	SEQ ID NO: 26
	9656364	tolA	SEQ ID NO: 27
Eukaryotes			
LM	1743289	hydrophilic surface protein 2	SEQ ID NO: 28
	468328	hydrophilic surface protein	SEQ ID NO: 29
PF	3845179	predicted integral membrane protein	SEQ ID NO: 30
	4493889	circumsporozite protein	SEQ ID NO: 31

Table 2: list of hypothetical proteins

Organism	GI Number	SEQ ID NO	GI Number	SEQ ID NO
Eubacteria				
BB	2688482	SEQ ID NO: 32	2688343	SEQ ID NO: 37
	2688046	SEQ ID NO: 33	2688447	SEQ ID NO: 38
	2688045	SEQ ID NO: 34	2688540	SEQ ID NO: 39
	2688103	SEQ ID NO: 35	2688768	SEQ ID NO: 40
	2688333	SEQ ID NO: 36	2688793	SEQ ID NO: 41
CJ	6967728	SEQ ID NO: 42	6968409	SEQ ID NO: 46
	6967819	SEQ ID NO: 43	6968423	SEQ ID NO: 47
	6968034	SEQ ID NO: 44	6968200	SEQ ID NO: 48
	6968265	SEQ ID NO: 45		
CP	4377009	SEQ ID NO: 49	4377196	SEQ ID NO: 54
	4377120	SEQ ID NO: 50	4376483	SEQ ID NO: 55
	4377121	SEQ ID NO: 51	4376770	SEQ ID NO: 56
	4377216	SEQ ID NO: 52	4376779	SEQ ID NO: 57
	4376866	SEQ ID NO: 53	4376756	SEQ ID NO: 58

CT	3328515	SEQ ID NO: 59	3329121	SEQ ID NO: 61
	3329021	SEQ ID NO: 60		
HI	1574537	SEQ ID NO: 62	1574799	SEQ ID NO: 65
	1574414	SEQ ID NO: 63	3212225	SEQ ID NO: 66
	1574625	SEQ ID NO: 64	1574607	SEQ ID NO: 67
HP	2313229	SEQ ID NO: 68	2313894	SEQ ID NO: 71
	2313552	SEQ ID NO: 69	2314686	SEQ ID NO: 72
	2313684	SEQ ID NO: 70		
MG	1045905	SEQ ID NO: 73	1045811	SEQ ID NO: 74
MP	1674046	SEQ ID NO: 75	1674374	SEQ ID NO: 78
	1673719	SEQ ID NO: 76	1673775	SEQ ID NO: 79
	1673772	SEQ ID NO: 77		
MTUB	2113965	SEQ ID NO: 80	2909499	SEQ ID NO: 82
	2117265	SEQ ID NO: 81		
NM	7225315	SEQ ID NO: 83	7227030	SEQ ID NO: 86
	7226708	SEQ ID NO: 84	7227104	SEQ ID NO: 87
	7226768	SEQ ID NO: 85	7226645	SEQ ID NO: 88
PAER	9947556	SEQ ID NO: 89	9948900	SEQ ID NO: 91
	9949353	SEQ ID NO: 90	9948180	SEQ ID NO: 92
RP	3860652	SEQ ID NO: 93	3860651	SEQ ID NO: 94
TP	3322751	SEQ ID NO: 95	3322546	SEQ ID NO: 96
VCHO	9654409	SEQ ID NO: 97	9657724	SEQ ID NO: 102
	9654544	SEQ ID NO: 98	9657931	SEQ ID NO: 103
	9654912	SEQ ID NO: 99	9658035	SEQ ID NO: 104
	9656707	SEQ ID NO: 100	9658254	SEQ ID NO: 105
	9657609	SEQ ID NO: 101	9656580	SEQ ID NO: 106
Eukaryotes				
Pathogens				
PF	3845248	SEQ ID NO: 107	4493994	SEQ ID NO: 109
	3845292	SEQ ID NO: 108	4494004	SEQ ID NO: 110

LM	6996498	SEQ ID NO: 111	6562665	SEQ ID NO: 115
	6978417	SEQ ID NO: 112	6996509	SEQ ID NO: 116
	6899670	SEQ ID NO: 113	6433946	SEQ ID NO: 117
	6899664	SEQ ID NO: 114	5869911	SEQ ID NO: 118

Advantages

The method of the invention for identifying unique protein sequences useful as antiinfectives is *ab initio*. It does not need a teaching data set. These anti-infectives are useful as vaccine candidates, as diagnostics, and in studying drug responses. The method uses sequence attributes instead of sequence patterns. The invention is generally applicable to all genomes and is easy to implement in any setting. This approach results in reproducible results as the method does not depend on variable biochemical characterization of proteins. However, functional information from other systems is helpful in aiding testable predictions. The method of the invention can be used for newly sequenced pathogens to provide a set of candidates for rapid evaluation for the development of anti-infectives.